Petropoulos, F., Wang, X. and Disney, S.M., (2019), The inventory performance of forecasting methods: evidence from the M3-competition data. International Journal of Forecasting, 35 (1), 251-265. DOI: 10.1016/j.ijforecast.2018.01.004.

# The inventory performance of forecasting methods: evidence from the M3-competition data

Fotios Petropoulos[a,*], Xun Wang[b], Stephen M. Disney[b]

[a] *School of Management, University of Bath, UK*
[b] *Cardiff Business School, Cardiff University, UK*

**Abstract**

Forecasting competitions have been a major drive not only for improving the performance of forecasting methods but also for the development of new forecasting approaches. Despite the tremendous value and impact of these competitions, they suffer from the limitation is that performance is measured only in terms of forecast accuracy and bias, lacking utility metrics. Using the monthly industry series of the M3-competition, we empirically explore the inventory performance of widely used forecasting techniques, including exponential smoothing, ARIMA models, Theta method and approaches based on multiple temporal aggregation. We employ a rolling simulation approach and analyse the results for the order-up-to policy under various lead times. We find that methods based on combinations result in superior inventory performance, while Naïve, Holt and Holt-Winters perform poorly.

*Keywords:* forecasting, inventory, evaluation, utility metrics, bullwhip effect

## 1. Introduction

Accurate forecasting is claimed to be important for business operations. However, one should not only focus on minimizing the forecast error per se but maximising the utility of the forecasts. Forecasting competitions have increased our understanding with regards to the relative forecasting performance of different methods and approaches. The M and M3 forecasting competitions are the most impactful to-date, involving more than 4,000 time series combined with an evaluation of the forecasts produced by more than 30 different methods. However, little is known about the utility of the forecasting methods in inventory management settings.

The main result from the M competitions (Makridakis et al., 1982; Makridakis and Hibon, 2000) is that simple forecasting methods, such as exponential smoothing, can perform as good as, if not better than, more complex ones, such as the Box-Jenkins method. However

*Correspondence: Fotios Petropoulos, School of Management, School of Management, University of Bath, Claverton Down, Bath, BA2 7AY, UK.

*Email addresses:* f.petropoulos@bath.ac.uk (Fotios Petropoulos), WangX46@cardiff.ac.uk (Xun Wang), DisneySM@cardiff.ac.uk (Stephen M. Disney)

Petropoulos, F., Wang, X. and Disney, S.M., (2019), The inventory performance of forecasting methods: evidence from the M3-competition data. International Journal of Forecasting, 35 (1), 251-265. DOI: 10.1016/j.ijforecast.2018.01.004.

the ranking of the different methods depends on the selected error measure. It has been shown that forecast combinations usually outperform individual forecasts and forecasting performance decreases as the forecasting horizon lengthens.

For inventory forecasting, practitioners care about the variance of the forecast errors over the lead-time and review period (as these factors lead directly to inventory holding requirements) but also the variance of the forecasts (as this leads to the bullwhip effect, Wang and Disney, 2016). When demand forecasts are used for replenishment decisions, it is the resulting inventory performance that ultimately determines the suitability of a forecasting method. However, in the M and M3 competitions, performance of the various methods is measured only in terms of forecasting performance. We assert that forecasts are created to aid business decisions and those forecasts should be judged based on their economic consequences. If a more stable, but less accurate forecast results in less cost for the business, that is the forecast that should be used and this calls for an extension on evaluation based on the inventory performance. In particular, there are three aspects of inventory performance that are especially important to everyday operations: (holding and backlog) cost, amplification (of inventory and replenishment orders) and customer service level.

Our contributions are as follows,

- We extend the results of the M3-competition to include evaluations on inventory performance, using the monthly industry series of the M3-competition.
- We empirically explore the utility of widely used forecasting methods, such as exponential smoothing, and other methods originally evaluated in M3-competition, such as ARIMA and the Theta method. We also extend to methods that have been proposed after the M3-Competition took place, such as the automatic selection of the best exponential smoothing model (ETS), automatic ARIMA (AutoARIMA) and Multiple Aggregation Prediction Algorithm (MAPA) which have been shown to perform well in terms of forecasting performance measures.
- We employ a rolling simulation approach and analyse the performance of the forecasting mechanism when used inside the order-up-to policy (a commonly adopted replenishment policy in practice) with various lead times and target service levels.
- We propose a single metric for inventory performance that incorporates first-order (piecewise linear cost functions), second-order (quadratic cost functions) and service level, based on which the forecast methods can be compared and contrasted.

The remainder of the paper is organized as follows. Section 2 provides a short review of existing studies on the inventory performance of forecasting methods. Section 3 describes the empirical data, the forecasting methods and the inventory system model that were adopted in this research. Section 4 defines the forecasting and utility measures considered. Section 5 presents the empirical results, and Section 6 concludes.

## 2. Research background

In this section we present an overview of the literature that has dealt so far with assessing the inventory performance of various forecasting methods.

Petropoulos, F., Wang, X. and Disney, S.M., (2019), The inventory performance of forecasting methods: evidence from the M3-competition data. International Journal of Forecasting, 35 (1), 251-265. DOI: 10.1016/j.ijforecast.2018.01.004.

In the slow-moving demand context, Sani and Kingsman (1997) were the first to assess the performance of various estimators, namely simple moving averages (SMA), simple exponential smoothing (SES) and Croston's method (Croston, 1972), under different inventory control systems. They did so by measuring the average regret on annual cost and the average service regret by using each forecasting method. Eaves and Kingsman (2004) additionally considered the Syntetos-Boylan Approximation (SBA, Syntetos and Boylan, 2005). They suggested that SBA can result in reduction of the inventory holding volumes compared to other intermittent demand estimators. Syntetos and Boylan (2006) performed a similar analysis considering both the stock volume and the realised customer service level. Boylan et al. (2008) extended this work to the case of classifying stock keeping units (SKUs) based on their demand characteristics (Syntetos et al., 2005). In a latter study, Babai et al. (2010) additionally considered the inventory performance of simple exponential smoothing (SES) along with Croston's method and SBA, while focusing on holding and backorder trade-offs.

The utility performance of parametric methods for forecasting intermittent demand has also been compared against bootstrapping methods (Teunter and Duncan, 2009; Syntetos et al., 2015) and neural method techniques (Kourentzes, 2013), with interesting insights emerging from the inventory versus realised service levels and holding versus backlog volume trade-off curves. Kourentzes (2014) proposed and demonstrated the effectiveness (in terms of both traditional error measures and inventory performance) of two new cost functions for parameter optimization of intermittent demand methods.

Fewer studies have empirically examined the inventory performance of forecasting methods for fast-moving demand series. The first attempt, to the best of our knowledge, is by Gardner (1990). He analysed the SKUs classified as important (Class A) in a military inventory system. The trade-off curves of inventory investment versus service level (in terms of number of days to fulfill a backorder) showed the superiority of damped exponential smoothing over SES, Naïve and Holt's methods. A later study (Acar and Gardner, 2012) reached the same results. Snyder et al. (2002) focused on estimating the aggregate (lead-time) demand using exponential smoothing methods and evaluated the performance on the fill rate.

A large empirical study of 28,000 SKUs was undertaken by Strijbosch et al. (2011). However, they compared the performance of only two simple forecasting methods: SMA and SES. Liao and Chang (2010) considered a wider pool of methods that includes the most commonly used ones from the exponential smoothing family (SES, Holt, Damped). However, the study was limited to a small number of time series. The inventory implications of forecast combination of a wide range of forecasting methods were examined by Barrow and Kourentzes (2016), suggesting that lower safety stocks are required compared to the base models. Lastly, Ali et al. (2012) examined the effects of information sharing in terms of forecasting on inventory savings (inventory holding volumes and inventory costs) under three forecasting methods (the optimal methods corresponding to AR(1), MA(1) and ARMA(1,1) series).

Some studies have examined the inventory implications of judgmentally revising statistical forecasts. Syntetos et al. (2009) showed that the benefits of judgmental interventions on forecast accuracy are also evident in terms of stock control performance (stock and ser-

Petropoulos, F., Wang, X. and Disney, S.M., (2019), The inventory performance of forecasting methods: evidence from the M3-competition data. International Journal of Forecasting, 35 (1), 251-265. DOI: 10.1016/j.ijforecast.2018.01.004.

vice levels). Syntetos et al. (2010) also demonstrated that, when fast demand is considered, small improvements in terms of accuracy as a result of judgmental revision of the statistical output could mean large gains in terms of inventory reductions. Wang and Petropoulos (2016) considered the cases of combination and forecast selection. They found that simple combinations of statistical and judgmental forecasts, or the appropriate selection between them, can lead to statistically significant reductions in the total inventory cost (consisting of holding cost and backlog cost) and variance of the inventory and the orders (which can be translated to reduction of the bullwhip effect, Wang and Disney, 2016), while maintaining similar achieved service levels compared to either statistical or judgmental forecasts.

This short literature review suggests that there exist many open research questions regarding the inventory performance of forecasting methods for fast-moving demand series:

- How does Holt-Winters method, arguably the most widely used method for fast-moving seasonal data, perform in terms of utility?
- What is the relative inventory performance of exponential smoothing over ARIMA?
- What is the inventory performance of the Theta method (Assimakopoulos and Nikolopoulos, 2000), the M3-Competition winner?
- Building on the research by Barrow and Kourentzes (2016), how do combinations of forecasting methods perform relative to the base models in terms of inventory volumes and realized service levels?
- While a few papers (Babai et al., 2012; Petropoulos et al., 2016) have studied the inventory implications of temporal non-overlapping aggregation (Nikolopoulos et al., 2011), a question that still remains is what is the inventory performance of multiple temporal aggregation (Kourentzes et al., 2014)?

## 3. Experimental design

### *3.1. Empirical data*

The empirical data for this study is the "monthly industry" subset of the M3-competition data (Makridakis and Hibon, 2000). This allows us to investigate the utility implications of forecasting methods in an operations related context, whilst retain the results of the original M3 competition as the accuracy benchmark.

We consider the monthly data, where utility will most probably relate to inventory. While many real-life inventory management settings operate on a daily or weekly basis, following the popular lean production philosophy which advocates as short a cycle as possible (Hedenstierna and Disney, 2016), there are still supply chains that operate on a monthly basis; usually supply chains with long lead-times and high capacity costs. We focus on the monthly frequency time series within the M3 dataset as data with a weekly frequency is not available. Such frequency corresponds to the inventory review period. Hence, adopting monthly data implies that the inventory levels are reviewed and ordering decisions are made each month. Since the focus of this paper is not setting the optimal review period, we refer to Silver and Robb (2008) or Hedenstierna and Disney (2016) for more information on this issue.

Petropoulos, F., Wang, X. and Disney, S.M., (2019), The inventory performance of forecasting methods: evidence from the M3-competition data. International Journal of Forecasting, 35 (1), 251-265. DOI: 10.1016/j.ijforecast.2018.01.004.

We exclude M3 time series falling under the categories "micro", "macro", "finance", "demographic" and "other" as these are not directly relevant to the nature of the problem investigated here nor could we reasonably assume that such data is part of an inventory system. Thus, we only consider the industry data with monthly frequency. After examining the description of the "monthly industry" subset in M3 data, we have found that, out of 334 time series, 256 directly corresponds to sales, shipment or production. These can be safely assumed to resemble the demand to be forecasted in the inventory system. For instance, the production data can be easily transformed to the demand or consumption of spare parts. In general, the network structure of supply chains allows us to translate non-demand data to the demand or consumption of some other products. Although there are some cases representing services rather than physical stock keeping units, we opt for not excluding these time series so that the M3-industry subset remains intact and the results provided from this work become comparable with the ones from the original research (Makridakis and Hibon, 2000). Insights derived from analysing just the 256 series that directly corresponds to sales, shipment or production are in-line with the results presented in this study.

Monthly or even seasonal data at the industry level has been widely applied in inventory and bullwhip analysis (Cachon et al., 2007). More importantly, the microscopic inventory system model have been used to model and analyse such data (Blinder and Maccini, 1991). This is an acceptable research norm for the inventory theorists, particularly when business level data with a finer granularity is not available. From the technical perspective, data from the industry level can be seen as an aggregation of business level data. Since it is difficult to track the aggregation policies in use, and hence the effect of aggregation on data structure and characteristics, we believe that it is safe for us to focus on the characteristics of the data.

The subset under investigation consists of 334 monthly time series of varying lengths (at least 96 and at most 144 observations), with the vast majority of the series recording at least 11 years of history. The data set consists of both stationary and non-stationary time series. We employ a rolling origin forecasting approach that allows for the inventory simulation discussed in Section 3.3. In more detail, we initially set the in-sample to hold the first 36 observations of the data and we produce forecasts for the next year (for the next 12 months, periods 37 to 48). We then include the $37^{\text{th}}$ observation in the in-sample and we once again produce forecasts for the next 12 months (periods 38 to 49). A forecast horizon of 12 periods corresponds to the maximum lead time that we consider in this research, which includes transportation lead-time and review period. We repeat this procedure until we reach the end of the data. For the shortest of the available series, this procedure is repeated 60 times (the first origin is the $36^{\text{th}}$ observation; the last origin is the $95^{\text{th}}$ observation from which only the first-step-ahead forecast can be evaluated).

### *3.2. Forecasting methods*

In this research we cannot make use of the publicly available forecasts submitted by the original participants of the M3-competition, as these forecasts are produced from a single forecast origin that would render the inventory evaluation practically impossible. However, we include methods that featured in the M3-competition that are freely available

Petropoulos, F., Wang, X. and Disney, S.M., (2019), The inventory performance of forecasting methods: evidence from the M3-competition data. International Journal of Forecasting, 35 (1), 251-265. DOI: 10.1016/j.ijforecast.2018.01.004.

(i.e. they do not require proprietary software), widely used in practice, and/or achieved good performance.

Apart from the simple Naïve (random walk) method, where the forecast for the next period is equal to the latest recorded actual, we consider several widely used methods coming from the exponential smoothing family, namely simple exponential smoothing (SES), Holt's exponential smoothing (Holt), Holt-Winters exponential smoothing (Holt-Winters) and Damped trend exponential smoothing (Damped).

We also include in our pool the standard version of the Theta method, as proposed by Assimakopoulos and Nikolopoulos (2000). The Theta method is a data-decomposition approach that attempts to separate the short term movements from the long-term trend of the data. Hyndman and Billah (2003) and Fiorucci et al. (2016) showed that the standard version of Theta method that was applied at the M3-Competition is mathematically equivalent to SES with drift. Given that the Theta method outperformed all other methods in the original results of the M3-competition, it would be interesting to see if its superior performance holds when considering inventory implications rather than traditional error measures used in the forecasting literature.

For the methods outlined so far and that do not model seasonality explicitly, we consider, in line with the original set-up of the M3-competition, a three-step forecasting procedure: seasonal adjustment (if data are seasonal), forecasting, re-seasonalization. In more detail, a seasonal test is applied using the in-sample data (that is used for fitting the different methods). The seasonal test applied in this study is described in detail by Fiorucci et al. (2016) and is based on the auto-correlation function with lag equal to 12 periods, corresponding to the seasonal cycle for the monthly frequency. We opt for a 90% confidence level for the seasonality test. If a series is identified as seasonal, then a multiplicative classical decomposition is applied to calculated the seasonal indices. The in-sample data are divided by the respective seasonal indices towards the calculation of the seasonally-adjusted in-sample data; at the same time, the hold-out sample remains intact. Then, the non-seasonal forecasts are produced using the seasonally-adjusted in-sample observations. Finally, these forecasts are re-seasonalised using the seasonal indices derived by the classical decomposition. The final seasonal forecasts are then compared with the hold-out sample. This procedure is repeated for each forecast origin. If data are not identified as seasonal, then forecasting is performed on the original data. Lastly, it should be noted that a 12-month cycle encapsulates commonly observed shorter business cycles, such as quarterly ones linked with financial reporting activities.

We additionally consider forecasting approaches proposed after the M3-competition and have shown to perform well. These include the algorithms proposed by Hyndman and Khandakar (2008) referring to automatic model selection for the exponential smoothing and the ARIMA families of models. Hereafter these approaches will be referred to as ETS and AutoARIMA respectively. The selection is based on information criteria with the default option of these automatic approaches being the corrected Akakie Information Criterion for finite sample sizes (AICc). We choose to add these approaches in our study as these are implemented as functions of the very popular *forecast* package of the `R` statistical software

Petropoulos, F., Wang, X. and Disney, S.M., (2019), The inventory performance of forecasting methods: evidence from the M3-competition data. International Journal of Forecasting, 35 (1), 251-265. DOI: 10.1016/j.ijforecast.2018.01.004.

and are nowadays considered as the benchmarks in automatic time series forecasting.[1]

We also investigate the performance of a recently proposed combination approach, the Multiple Aggregation Prediction Algorithm (MAPA, Kourentzes et al., 2014), which is based on the temporal transformation of the original time series. MAPA suggests that forecasts are not produced using just the original (monthly in this case) frequency, but also higher non-overlapping temporal aggregation levels (lower frequencies), such as bi-monthly, quarterly, up to the yearly frequency. The advantage is that different time series features (trend and seasonality) will be enhanced and/or smoothed out at different aggregation levels, rendering their extrapolation more robust. Consequently, the estimates of the states (level, trend and seasonality) at the different levels are combined towards the final forecast.

The inclusion of MAPA is based on the increased popularity of the temporal aggregation since 2011 coupled with their increased forecast accuracy, especially if one focuses on the longer horizons. Additionally, approaches based on the combination of various temporal aggregation levels are of managerial importance, as the produced forecasts are aligned to operational, tactical and strategic decision making purposes.

Table 1 summarises the methods and approaches implemented in this research. In line with the reproducibility agenda (Boylan et al., 2015; Boylan, 2016) this table also details the `R` packages and functions that were used to produce the forecasts.

Table 1: Forecasting methods and approaches used in this study.

| Method | `R` package | Function |
|---|---|---|
| Naïve | *forecast* 6.2 | `naive(...)` |
| SES | *forecast* 6.2 | `ets(..., model="ANN")` |
| Holt | *forecast* 6.2 | `ets(..., model="AAN", damped=FALSE)` |
| Holt-Winters | *forecast* 6.2 | `ets(..., model="MAM", damped=FALSE)` |
| Damped | *forecast* 6.2 | `ets(..., model="AAN", damped=TRUE)` |
| Theta | *forecTheta* 2.1 | `stheta(...)` |
| ETS | *forecast* 6.2 | `ets(...)` |
| AutoARIMA | *forecast* 6.2 | `auto.arima(...)` |
| MAPA | *MAPA* 1.9 | `mapasimple(...)` |

Apart from single methods and approaches, the performance of combinations, which have been shown to perform well in the M3-competition, is explored. More specifically, we consider the simple (equal-weight) combination of SES, Holt and Damped methods (denoted as SHD) and the simple combination of the ETS and AutoARIMA approaches.

We would like to note that the scope of this study is to compare the various methods and approaches when these are applied "out-of-the-box", i.e. using implementations that can

[1]The *forecast* package is currently ranked within the top 1% (across all R packages available in CRAN repository) in terms of monthly downloads, with an average of 56,000 downloads per month. Also, the study that proposed ETS and autoARIMA automatic forecasting approaches (Hyndman and Khandakar, 2008) has been cited more than 800 times so far based on Google Scholar.

Petropoulos, F., Wang, X. and Disney, S.M., (2019), The inventory performance of forecasting methods: evidence from the M3-competition data. International Journal of Forecasting, 35 (1), 251-265. DOI: 10.1016/j.ijforecast.2018.01.004.

be found within the solutions of popular forecasting software providers. For instance, the parameters for the exponential smoothing methods have been estimated by minimising the one-step-ahead mean squared error. We do not investigate what are the potential gains of matching the optimisation objective with a particular utility objective, such as minimisation of the holding cost or maximisation of the achieved service level; we consider this as a separate study.

### *3.3. Inventory system*

The evaluation method proposed does not need the specification of an inventory replenishment policy, as long as it takes demand and forecast as input. However, for the sake of simulation studies, here we adopt the order-up-to policy as the inventory replenishment policy. This policy has been frequently applied to industrial practices as well as academic research, due to its simple structure and ease of implementation (Lee et al., 1997). It is also the optimal linear policy for minimizing inventory related costs when there is no fixed ordering cost (Hosoda and Disney, 2006).

The policy is defined as follows:

$$o_t = \hat{D}_t^L + ss_t - ip_t, \tag{1}$$

where $o_t$ is the determined ordering quantity in period $t$, $\hat{D}_t^L$ the forecast of lead-time demand and $L$ is the lead-time; $ss_t$ the safety stock, and $ip_t$ the inventory position at the end of period $t$ respectively. Note that the term $L$ includes the transportation lead-time and a one sequence of events delay.

The lead-time demand and the forecast of lead-time demand are calculated simply by summing single-period demand and forecasts within the lead-time, i.e.,

$$D_t^L = \sum_{k=1}^{L} d_{t+k}, \tag{2}$$

$$\hat{D}_t^L = \sum_{k=1}^{L} \hat{d}_{t,t+k}, \tag{3}$$

where $d_t$ is the demand (consumption) in period $t$, and $\hat{d}_{t,t+k}$ is the forecast of demand in period $t+k$ made at period $t$. The single-period forecasts $\hat{d}_{t,t+k}$ are generated by the forecasting methods discussed in Section 3.2.

The safety stock is updated iteratively in a Newsvendor fashion. That is, it is determined by the desired service level (availability), $\alpha_s$, and the standard deviation of past lead-time demand forecast errors,

$$ss_t = \Phi^{-1}(\alpha_s)\sigma_e, \tag{4}$$

where $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function of the standard normal distribution, and $\sigma_e$ is the standard deviation of past forecast errors, calculated from the first period up to period $t$. The adoption of order-up-to policy is needed here to ensure that we

can use the standard deviation of forecast errors ($\sigma_e$) to replace the standard deviation of inventory levels ($\sigma_i$). We acknowledge that often exponential smoothing is used in practice to update $\sigma_e$ (Chen et al., 2000).

Lastly, the inventory position is updated every period following the linear assumption,

$$ip_t = ip_{t-1} + o_{t-1} - d_t. \tag{5}$$

The linear assumption implies that unmet demand can be backlogged and the capacity of supply is infinite. We keep these assumptions for the simplification of the inventory system model, as relaxing them would potentially hinder our focus on the impact of forecasting methods. The inventory position can further be broken down to net inventory level and work-in-progress level (ordered placed but not yet received), $ip_t = i_t + w_t$, where

$$i_t = i_{t-1} + o_{t-L} - d_t,$$

$$w_t = w_{t-1} + o_{t-1} - o_{t-L}.$$

## 4. Measuring performance

### *4.1. Forecasting performance*

In this study we consider one bias and two accuracy measures. The Mean Percentage Error ($MPE$) is used to measure bias. A single point forecast Percentage Error ($PE$) is defined as

$$PE = \frac{d_t - \hat{d}_t}{d_t} \times 100\%. \tag{6}$$

The $MPE$ can be calculated as the arithmetic mean of $PE$s across horizons, origins and time series.

To measure accuracy, we consider the symmetric Mean Absolute Percentage Error ($sMAPE$), which was also the main error measure in the M3-competition, as well as the Mean Absolute Scaled Error ($MASE$, Hyndman and Koehler, 2006). A symmetric Absolute Percentage Error ($sAPE$) and a Absolute Scaled Error ($ASE$) can be defined as

$$sAPE = \frac{\mid d_t - \hat{d}_t \mid}{\mid d_t \mid + \mid \hat{d}_t \mid} \times 200\%. \tag{7}$$

$$ASE = \frac{\mid d_t - \hat{d}_t \mid}{(n-m)^{-1} \sum_{i=m+1}^{n} \mid d_i - d_{i-m} \mid}, \tag{8}$$

where $m$ is the periodicity of the data (12 for monthly time series) and $n$ is the length of the in-sample. The $sMAPE$ and $MASE$ can be computed by appropriately calculating the arithmetic means of the $sAPE$s and $ASE$s respectively across horizons, origins and time series.

Petropoulos, F., Wang, X. and Disney, S.M., (2019), The inventory performance of forecasting methods: evidence from the M3-competition data. International Journal of Forecasting, 35 (1), 251-265. DOI: 10.1016/j.ijforecast.2018.01.004.

### *4.2. Inventory performance*

There are three categories of performance metrics typically involved in evaluating an inventory system: financial, operational, and service. The financial metrics look at the cost incurred in the system with regards to inventory holding, backlog, and ordering. On the other hand, due to the limited availability of financial data, sometimes only operational metrics, such as the order and inventory variance, are considered, as their calculation requires only quantity related information. Lastly, service metrics measure the percentage of periods that end with inventory available (availability) or the proportion of demand satisfied directly from the stock (fill rate, Disney et al., 2015).

From a modelling perspective, often a cost function has to be considered which dictates the relationship between financial cost and quantity. For instance, the linear cost function in the classic Economic Order Quantity (EOQ) model assumes that the total inventory holding costs are linearly increasing with the level of inventory-on-hand. The fixed ordering cost implies that the ordering cost is a step function of the order quantity. The metrics can then be categorized according to the type of cost function.

Linear (or piecewise-linear) cost functions are most commonly adopted in inventory control literature, which can be found in EOQ, Newsvendor and ($s$,$S$) models (where $s$ is the reorder point and $S$ is the order-up-to level) (Axsäter, 2006). It is assumed that the cost and quantity form a linear (or piecewise-linear) relationship. This assumption is deemed valid when the contribution of each item to the total cost remains constant. In this study, the total inventory-related cost is considered to be the summation over holding and backlog costs, which are linear to inventory-on-hand and backlog levels respectively:

$$TC = C_h + C_b = h\mathbb{E}\left[(i_t)^+\right] + b\mathbb{E}\left[(-i_t)^+\right], \tag{9}$$

where $h$ and $b$ are unit holding and backlog costs respectively. $(\cdot)^+ = \max(\cdot, 0)$ is the maximum function used to indicate whether the inventory is sufficient to satisfy customers' demand and $\mathbb{E}[\cdot]$ is the expectation operator. There is inventory-on-hand if $i_t > 0$, and a backlog if $i_t < 0$.

Another frequently used cost function is a quadratic one. The relationship between cost and quantity is assumed to be quadratic (i.e. of second-order). This is a special case of diseconomy of scale, where the increment of cost increases with quantity. The well-known bullwhip effect and inventory amplification, commonly defined as the variance of order quantity ($v_o$) and inventory level ($v_i$), can be seen as quadratic cost functions (Wang and Disney, 2016).

When the inventory distribution is normal and perfectly known and the safety stock is set as (4), the total cost is then

$$TC = \sqrt{v_i}(b+h)\phi\left[\Phi^{-1}\left(\frac{b}{b+h}\right)\right]. \tag{10}$$

in which there is a direct correlation between inventory variance and inventory cost. However it is important to note that in the current study, such relationship does not strictly follow (10) as the neither of the assumptions is satisfied.

Petropoulos, F., Wang, X. and Disney, S.M., (2019), The inventory performance of forecasting methods: evidence from the M3-competition data. International Journal of Forecasting, 35 (1), 251-265. DOI: 10.1016/j.ijforecast.2018.01.004.

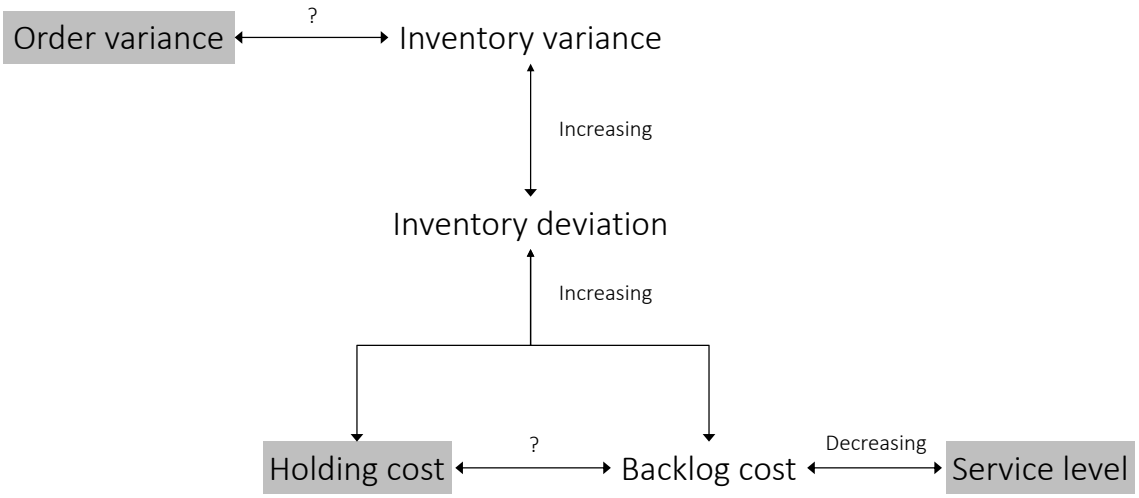


Figure 1: Relationship between the inventory performance measures. The highlighted measures contribute to the calculation of $RMS$ in Section 4.3.

The availability metric $\alpha$ (a.k.a. service level), measures the likelihood that a period ends with stock on-hand. It can be evaluated retrospectively with historical inventory data as the probability of non-negative inventory level:

$$\alpha = \Pr\{i_t \geqslant 0\}$$

The safety stock formula (4) is designed for normally distributed forecast errors, which is a common assumption in most operations management textbooks. This is not necessarily true for the empirical data-set that we have used in this paper. Also, the fitted distribution based on available data can never be identical with the real one. These lead to a deviation between the target and achieved service level, and a fact that the achieved service level cannot be estimated in the Newsvendor fashion.

It is obvious to see that the five metrics introduced above ($C_h$, $C_b$, $v_o$, $v_i$, and $\alpha$) are not independent with one another. For instance, as the service level $\alpha$ measures the probability of non-negative inventory, and backlog cost $C_b$ is determined by the level of negative inventory, it is safe to infer that a negative relationship exists between $\alpha$ and $C_b$. On the contrary, the relationship between $v_o$ and $v_i$ is not deterministic and dependent on various factors. There is research suggesting that trade-off exists between these two metrics, especially when one tries to adjust the inventory feedback in the ordering policy (Disney et al., 2004). Furthermore, $v_o$ and $v_i$ could change in the same direction with the influence of forecast errors (Wang and Petropoulos, 2016). Similarly, $C_h$ and $C_b$ do not have a monotonic relationship, as they are both affected by the mean and standard deviation of net inventory. Figure 1 shows the relationship between these metrics.

Petropoulos, F., Wang, X. and Disney, S.M., (2019), The inventory performance of forecasting methods: evidence from the M3-competition data. International Journal of Forecasting, 35 (1), 251-265. DOI: 10.1016/j.ijforecast.2018.01.004.

### *4.3. Holistic measure of inventory performance*

In most inventory scenarios, decisions will be based on more than one of the three inventory performance measures suggested in Section 4.2, due to trade-offs and interactions among these measures. Most existing papers use the sum of holding cost and backlog cost, overlooking the cost induced by ordering swings (Axsäter, 2006). There are attempts to adopt multiple criteria in past literature on inventory control. For instance, a weighted sum of order variance and inventory variance (Disney et al., 2004); trade-off curve between holding cost and service level (Babai et al., 2012); the costs of inventory holding and backlogs and production idling and overtime, (Hosoda and Disney, 2012). However, there are concerns over these cost function settings, e.g., that the weighting factors are difficult to set in practice, and that the forecasting methods are not easily ranked when multiple criteria are present.

Consequently, we combine the utility metrics mentioned in the previous subsection towards a single measure. To that end, we exclude the standard deviation of inventory $v_i$ and the backlog cost $C_b$ as to avoid correlations in measurement. The correlation is strong between $v_o$ and $v_i$ ($\rho = 0.686$) and between $C_b$ and $\alpha$ ($\rho = -0.516$). Another reason that we exclude $C_b$ is that the unit backlog cost is difficult to estimate in practice and businesses prefer to use service level to measure the probability of stockout. Thus, we end up with the variance of orders $v_o$, the holding cost $C_h$ and the service level $\alpha$. These three measures are widely used in industrial practice and they are highlighted in Figure 1 with grey shades. Moreover, the correlation between the three measures are low ($\rho < 0.25$).

As such, for the current multi-criteria optimisation problem we propose utilising the root mean square ($RMS$). Assuming two criteria, $A$ and $B$, for which we wish to minimise their value, $RMS$ can be written as

$$RMS = \sqrt{\frac{1}{2}\left(A^2 + B^2\right)}. \tag{11}$$

The adoption of $\mathcal{L}_2$ norm rather than $\mathcal{L}_1$ can be explained by the dis-economy of scale in operations. It is generally believed that a unit increase in inventory level when it is high brings more cost than when it is low. This is also true for shortage and order fluctuation. Therefore, the squared form of performance measure better reflects the true cost structure (Holt et al., 1960).

Equation (11) implies that equal weights are associated with $A$ and $B$. The equal weighting is derived from the notion that the decision maker gives equal emphasis on inventory holding, shortage and order fluctuation. We acknowledge that this is often not the case in practice, where planners focus more on one measure than the others. In Section 5.4 we will show through a sensitivity analysis that the weights do not significantly affect the final results.

Also, the units and scale of $A$ and $B$ can be different. In our case, holding cost is measured in dollars, standard deviation by product quantity and service level by percentage. We suggest normalising each criterion by dividing its value with the mean value of the same

Petropoulos, F., Wang, X. and Disney, S.M., (2019), The inventory performance of forecasting methods: evidence from the M3-competition data. International Journal of Forecasting, 35 (1), 251-265. DOI: 10.1016/j.ijforecast.2018.01.004.

criterion across all cases considered (in this study, across all forecasting methods), or:

$$RMS = \sqrt{\frac{1}{2}\left[\left(\frac{A}{\overline{A}}\right)^2 + \left(\frac{B}{\overline{B}}\right)^2\right]}. \tag{12}$$

Equation (12) suggests that the relative improvement/deterioration of each criterion over the average case is fed into the calculation of $RMS$, which is analogous to the introduction of percentage error in forecasting literature.

Considering the three inventory performance indicators identified in the previous subsection ($C_h$, $v_o$ and $\alpha$) and accounting for the fact that achieved service level has to be inverted, the $RMS$ for our problem is expressed as

$$RMS = \sqrt{\frac{1}{3}\left[\left(\frac{C_h}{\overline{C_h}}\right)^2 + \left(\frac{v_o}{\overline{v_o}}\right)^2 + \left(\frac{\overline{\alpha}}{\alpha}\right)^2\right]}. \tag{13}$$

Note that $C_h$ and $v_o$ are to be minimized, whereas $\alpha$ is to be maximized. Thus we take the inverse of the last term in (13). Consequently, the forecasting method with lower $RMS$ is considered to be superior to those with higher $RMS$.

## 5. Results

In this section we will present the empirical results of the study. First, we will show the forecasting performance of the various methods considered (Section 5.1). Then, inventory performance is presented in terms of order variance versus inventory variance (Section 5.2) and holding cost versus achieved service level (otherwise known as an efficient frontier, Section 5.3). Lastly, we combine the various inventory performance metrics using $RMS$ (Section 5.4).

### *5.1. Results on forecasting performance*

Table 2 presents the forecasting performance results of the various methods and combinations considered. Accuracy is measured in terms of $sMAPE$ and $MASE$, while bias is measured in terms of $MPE$. Results in this table are presented for lead-time equal to 12 periods, as a long lead-time amplifies the differences in performance which allows us for better observation. We do note however that a 12-month lead-time seldom exist in real-life supply chains. Also, we have found that the insights for the other lead times (1,3 and 6) are similar.

We observe that there is a great deal of disagreement between the bias and the accuracy measures. Naïve is ranked 1$^{st}$ in terms of bias, but is between the bottom-three methods in terms of both $sMAPE$ and $MASE$. On the other hand, MAPA and ETS-AutoARIMA perform badly in terms of bias (8$^{th}$ and 9$^{th}$ respectively), but are ranked 1$^{st}$ and 2$^{nd}$ in terms of accuracy. In any case, all methods over-forecast.

Table 2: Forecasting performance of the various methods and combinations ($L = 12$). Ranks are presented in brackets.

| Method | $MPE$ (%) | $sMAPE$ (%) | $MASE$ |
|---|---|---|---|
| Naïve | −2.530 (1) | 12.824 (9) | 0.945 (9) |
| SES | −2.718 (3) | 11.759 (3) | 0.865 (5) |
| Holt | −3.059 (7) | 13.544 (11) | 0.961 (11) |
| Damped | −2.769 (4) | 12.140 (6) | 0.883 (7) |
| Holt-Winters | −3.728 (11) | 13.444 (10) | 0.960 (10) |
| Theta | −2.845 (5) | 11.923 (5) | 0.862 (4) |
| ETS | −2.653 (2) | 11.819 (4) | 0.851 (3) |
| AutoARIMA | −3.661 (10) | 12.211 (7) | 0.880 (6) |
| MAPA | −3.108 (8) | 11.322 (1) | 0.832 (2) |
| SHD | −2.849 (6) | 12.248 (8) | 0.884 (8) |
| ETS-AutoARIMA | −3.157 (9) | 11.499 (2) | 0.827 (1) |

Damped exponential smoothing confirms its title as a benchmark method for time series forecasting, achieving a median rank across all methods, being slightly better in terms of bias. SES performs exceptionally well in this subset of M3-Competition, being ranked 3$^{\text{rd}}$ in terms of $MPE$ and $sMAPE$ and 5$^{\text{th}}$ in terms of $MASE$. The performance of the Theta method confirms the results of M3-Competition, ranked second after SES amongst the methods originally participated in the large empirical exercise by Makridakis and Hibon (2000). However, it is outperformed by more recently proposed approaches, such as MAPA and ETS.

We should mention that Holt and Holt-Winters are the two methods with the worst performance amongst all contenders. This is not a surprising as these methods did not perform well in the original competition either.

In Figure 2, we also present the results from *multiple comparisons from the best* (MCB) test, as applied by Koning et al. (2005). Essentially, MCB tests if the average (across time series) rank of each method is significantly different from that of other methods. If the intervals of two methods do not overlap, this indicates statistically different performance.

The panels for $sMAPE$ and $MASE$ generally agree with the results presented in Table 2. MAPA and ETS-AutoARIMA are the top performers. ETS, SES and Theta follow without being significantly different. However, the last six methods' ranked performance (Damped, SHD, AutoARIMA, Holt-Winters, Naïve and Holt) is significantly worse than that of both MAPA and ETS-AutoARIMA.

The results of the panel for $MPE$ of Figure 2 show a slightly different picture from the respective column of Table 2. $MPE$ in Table 2 is calculated as the arithmetic mean of the percentage errors. This means that large positive and large negative errors cancel out. In MCB we first consider the absolute values of the signed percentage errors, so that average ranks make sense. In any case, all methods, apart from MAPA, perform similarly (their rank intervals overlap) in terms of bias according to the MCB test.

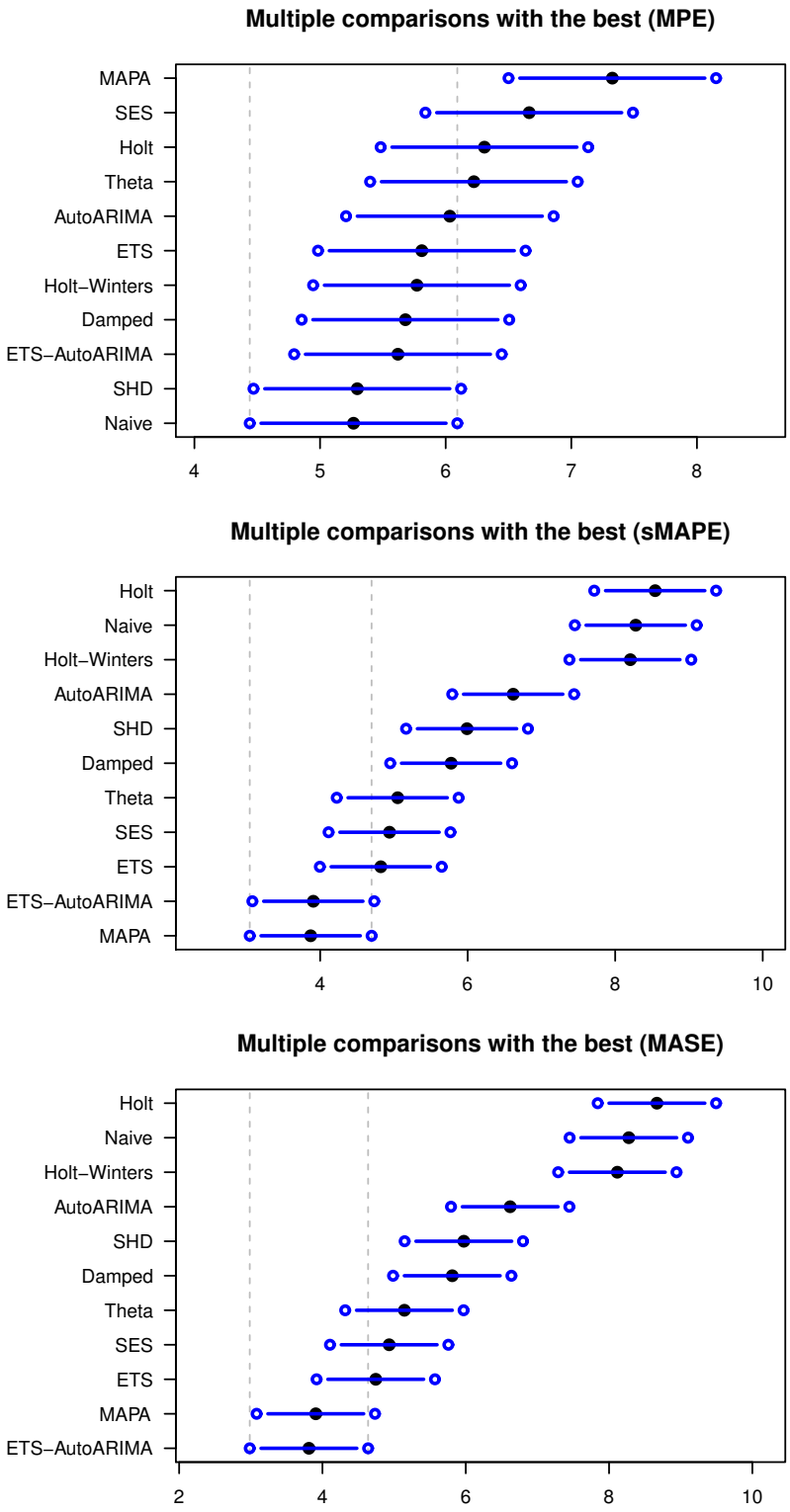


Figure 2: Multiple comparisons with the best based on the $MPE$, $sMAPE$ and $MASE$ (averaged per series, across horizons and origins).

Petropoulos, F., Wang, X. and Disney, S.M., (2019), The inventory performance of forecasting methods: evidence from the M3-competition data. International Journal of Forecasting, 35 (1), 251-265. DOI: 10.1016/j.ijforecast.2018.01.004.

### *5.2. Variance of orders and inventory*

Figure 3 shows order and inventory variance for each forecasting method as lead-time and service level varies. In all three service level cases (90, 95 and 99%), order and inventory variance increases with lead-time. One can compare the resulted order and inventory variance for the different methods when a specific lead time and target service level is assumed. Consider, for instance, that we are interested in lead-time equal to three periods and a target service level equal to 90%. The first panel of Figure 3 indicates that MAPA produces less order and inventory variance compared to all other methods.

In the order-up-to inventory system, in the long term under a constant safety stock, the variance of inventory equals to the variance of lead-time demand forecast error, $\text{var}(i) = \text{var}(e)$, $e_t = D_t^L - \hat{D}_t^L$, Hosoda and Disney (2012). On the other hand, the variance of orders is influenced by the amplitude of change (first-order difference) in lead-time demand forecast, and the relationship between such change and actual demand. This can be formally expressed by the following proposition:

**Proposition 1.** *If the safety stock is constant,* $v_o = \text{var}(\Delta\hat{D}^L) + \text{var}(d) + 2\text{cov}(\Delta\hat{D}^L, d)$, *where* $\Delta\hat{D}_t^L = \sum_{j=1}^{L} \hat{d}_{t,t+j} - \sum_{j=1}^{L} \hat{d}_{t,t-1+j}$ *is the change of lead-time forecast between periods* $t-1$ *and* $t$.

*Proof.* The proof is straightforward from (1) and (5), where one directly gets $o_t = \hat{D}_t^L - \hat{D}_{t-1}^L + d_t = \Delta\hat{D}_t^L + d_t$. □

Proposition 1 indicates that low order variance can be derived by (i) small fluctuation of the change in lead-time demand forecast; and (ii) negative correlation between demand and the change in lead-time demand forecast. Table 3 shows the ratios between $\text{var}(\Delta\hat{D}^L)$, $\text{cov}(\Delta\hat{D}^L, d)$, $\text{var}(e)$ and $\text{var}(d)$ for each forecasting method. $L = 12$ and $\alpha = 99\%$ is selected as an illustrative case.

Table 3: Variance and covariance of lead-time demand forecast and error when $L = 12$ and $\alpha = 99\%$. Ranks are presented in brackets.

| Method | $\frac{\text{var}(\Delta\hat{D}^L)}{\text{var}(d)}$ | $\frac{\text{cov}(\Delta\hat{D}^L,d)}{\text{var}(d)}$ | $\frac{\text{var}(e)}{\text{var}(d)}$ |
|---|---|---|---|
| Naïve | 87.656 (11) | 0.768 (11) | 109.478 (9) |
| SES | 22.690 (3) | 0.636 (6) | 81.006 (2) |
| Holt | 58.660 (10) | 0.701 (10) | 152.757 (11) |
| Damped | 32.975 (7) | 0.629 (5) | 97.421 (7) |
| Holt-Winters | 47.970 (9) | 0.601 (3) | 135.963 (10) |
| Theta | 29.269 (6) | 0.560 (2) | 94.025 (5) |
| ETS | 29.194 (5) | 0.668 (9) | 97.299 (6) |
| AutoARIMA | 21.915 (2) | 0.639 (7) | 84.074 (3) |
| MAPA | 11.280 (1) | 0.478 (1) | 71.898 (1) |
| SHD | 33.461 (8) | 0.655 (8) | 99.605 (8) |
| ETS-AutoARIMA | 24.903 (4) | 0.614 (4) | 87.348 (4) |

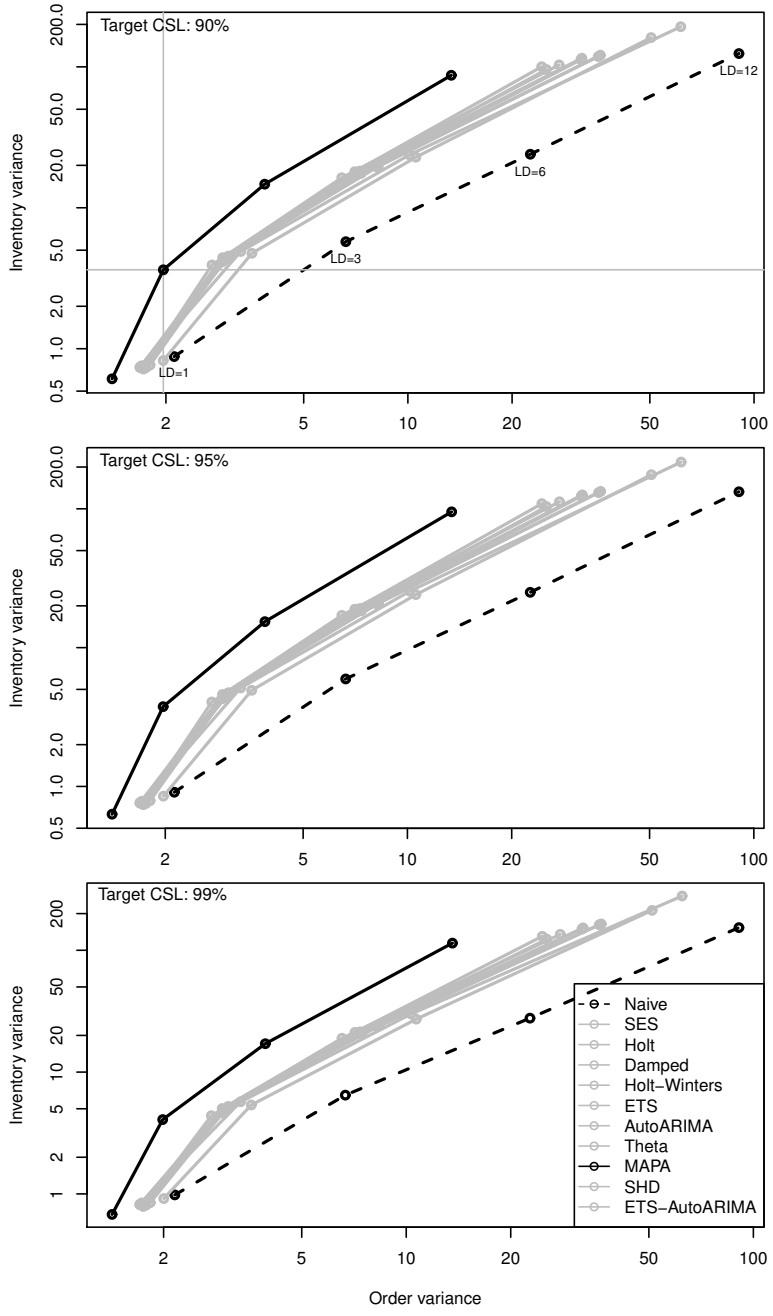


Figure 3: Order variance vs. inventory variance for various forecasting methods, target service levels and lead-times. Note that we have used log scales for clarity.

Petropoulos, F., Wang, X. and Disney, S.M., (2019), The inventory performance of forecasting methods: evidence from the M3-competition data. International Journal of Forecasting, 35 (1), 251-265. DOI: 10.1016/j.ijforecast.2018.01.004.

The values of $\text{cov}(\Delta\hat{D}^L, d)$ is negligible compared with $\text{var}(\Delta\hat{D}^L)$, therefore, we can conclude that the performance of order variance is mostly determined by the variance of change in lead-time demand forecast. The Naïve method generates the highest order change variance, which is the reason why it generates so much bullwhip. To make matters worse, the Naïve method also gives the highest variance of lead-time demand forecast errors, which solely determines the inventory variance.

In contrast, MAPA method provides both the lowest order and inventory variance, making it the strongest candidate in reducing inventory cost and bullwhip effect. The reason of this advantage, as shown in Table 3, is the fact that MAPA outputs both accurate (as measured in $MASE$) and smooth lead-time demand forecasts. The advantage of MAPA over other methods can also be seen for other lead-time and target service level scenarios, where the margins might be less significant.

Note that in this study we consider traditional parameter spaces for the forecasting methods. Li et al. (2014) showed that Damped method is capable of removing the bullwhip effect, but this requires some usually not recommended parameter settings that are not available in the R package.

### *5.3. Costs versus service levels*

Figure 4 presents the trade-off curves of holding cost (horizontal axis) versus achieved service level (vertical axis). Each curve shows the performance of a method for three different targeted service levels (90, 95, and 99%). The four panels correspond to the different lead times assumed in the analysis (1, 3, 6 and 12).

Each panel can be read as follows. Assuming a vertical line that corresponds to a specific holding cost (for example, 1000 for the first panel), then the various methods result in different achieved service levels; so we can relatively rank the methods based on the achieved service level. Similarly, assuming an horizontal line that corresponds to a constant achieved service level (for example, 90% in the first panel), then the different approaches can be relatively ranked in terms of realised holding cost. In other words, and given the trade-off relationship between holding cost and achieved service level (as the one gets worse, the other gets better), we should prefer the method depicted by curves that are closer to the top-left corner of each panel and we should avoid methods with by curves closer to the bottom-right corner. We present with solid black curves the methods that perform best for each lead time; with grey colour the methods that perform on average; and with dashed and dotted black curves the worst methods.

We observe that MAPA is among the top-performing methods for reducing inventory holding cost. This is especially the case for shorter lead times. AutoARIMA and the simple combination of ETS-AutoARIMA also performs well when lead-time is long ($L = 3, 6, 12$). On the other hand, Naïve, Holt and Holt-Winters are always among the worst performing methods when it comes to holding cost, especially Holt-Winters. All other methods constitute the gray middle class and seem to have indifferent performance in terms of holding cost. This means that there is little to distinguish between the performance of SES, Damped and Theta and the simple combination SHD.

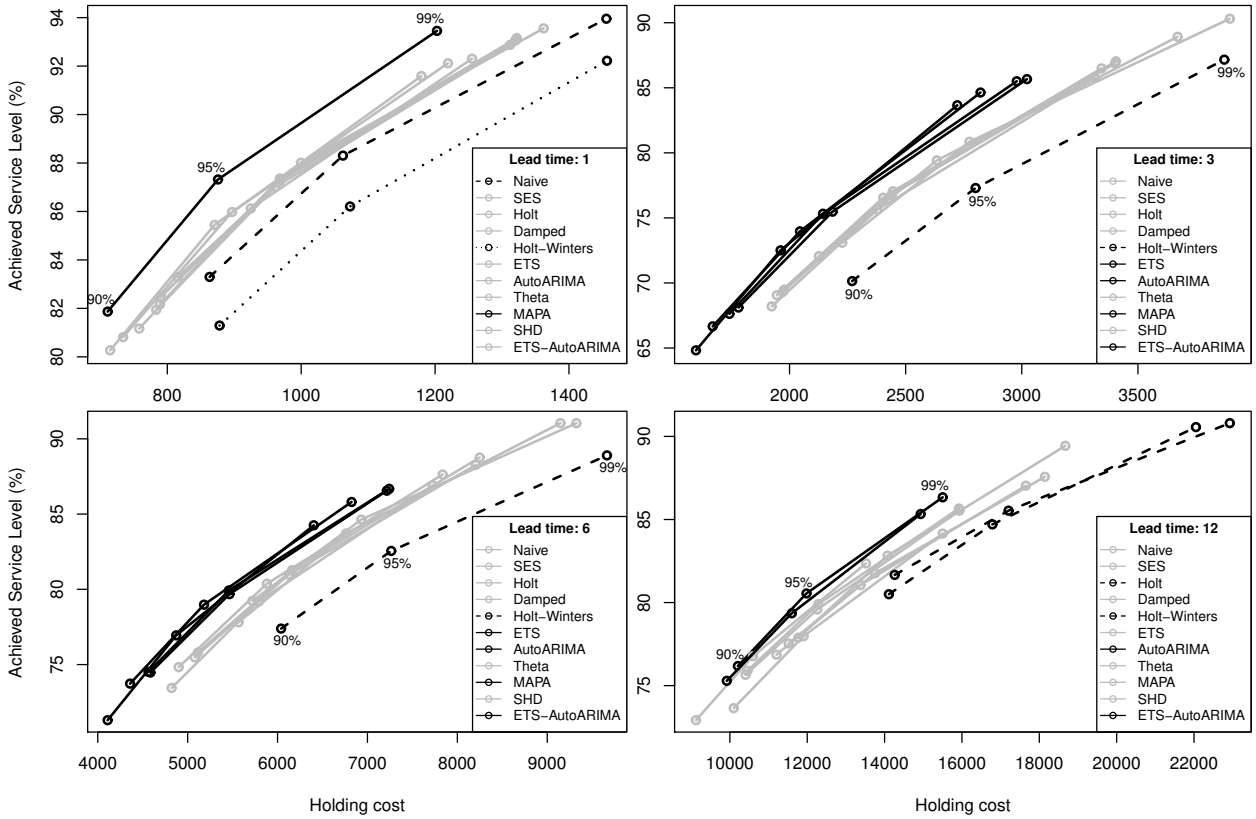


Figure 4: Trade-off curves of holding cost versus achieved service levels for different methods and lead times.

Petropoulos, F., Wang, X. and Disney, S.M., (2019), The inventory performance of forecasting methods: evidence from the M3-competition data. International Journal of Forecasting, 35 (1), 251-265. DOI: 10.1016/j.ijforecast.2018.01.004.

Another noteworthy observation is that the achieved service levels are lower than the targeted ones. While this can be observed for all lead times, it is especially the case when $L = 3$. This is because the distribution of forecast errors is most left-skewed when $L = 3$ ($-0.13$, negative skewness indicating long tail on the left), compared with other lead-time cases ($-0.07$ for $L = 1$, $-0.12$ for $L = 6$ and $-0.07$ for $L = 12$), averaged over all time series, forecasting methods and target service level. On the other hand, the safety stock calculation in the current research design assumes normally distributed inventory. Hence, the safety stock calculated under the normality assumption is not sufficient to maintain target service level with the actual inventory distribution.

The shape of inventory distribution is another reason for the differences in service level performance among forecasting methods, apart from the inventory fluctuation. Although MAPA is shown to be superior in maintaining a low standard deviation of inventory, it also gives the most right skewed inventory distribution ($-0.115$ in skewness, compared with 0.03 for Naïve). This result is linked with the bias measure of forecasting methods (first panel of Figure 2).

### *5.4. Combined performance and discussion*

Table 4 presents the inventory performance results in terms of the $RMS$ (equation 13) that was proposed in Section 4.3. The $RMS$ is averaged across all lead times (1 to 12) and all service levels (90, 95 and 99%). Column 2 presents the values of $RMS$ together with the ranks (in brackets) of the methods based on these values respectively. Column 3 presents the respective values for $MASE$ averaged across all horizons, 1 to 12.

Table 4: Average root mean square ($RMS$) values for the different methods (and ranks), contrasted with the $MASE$.

| Method | $RMS$ | $MASE$ |
|---|---|---|
| Naïve | 1.819 (11) | 0.945 (9) |
| SES | 0.976 (4) | 0.865 (5) |
| Holt | 1.134 (9) | 0.961 (11) |
| Damped | 0.986 (6) | 0.883 (7) |
| Holt-Winters | 1.183 (10) | 0.960 (10) |
| Theta | 0.952 (2) | 0.862 (4) |
| ETS | 0.991 (7) | 0.851 (3) |
| AutoARIMA | 1.004 (8) | 0.880 (6) |
| MAPA | 0.930 (1) | 0.832 (2) |
| SHD | 0.980 (5) | 0.884 (8) |
| ETS-AutoARIMA | 0.962 (3) | 0.827 (1) |

The first observation is that forecasting methods that are based on combinations (MAPA, Theta, SHD and ETS-AutoARIMA) perform very well in terms of their combined inventory performance. Note that the ranks of all these four methods improve compared to their respective ranks in terms of $MASE$. At the same time, the ranked performance of model

selection techniques, such as ETS and AutoARIMA, is decreased for $RMS$ compared to $MASE$. One should consider model selection frameworks directly linked to inventory performance, as Wang and Petropoulos (2016) show. In any case, the three forecasting methods that show the worst forecasting performance (Naïve, Holt and Holt-Winters), continue to perform poorly in terms of inventory.

Another noteworthy observation is that methods and approaches that have been introduced more recently (such as MAPA, Theta and the combination of ETS-AutoARIMA) and have proved to perform well in terms of accuracy, they also do well in terms of inventory performance.

Over all metrics of inventory performance, MAPA is ranked first with its ranked performance being significantly different than that of all other methods (see Figure 5 for the MCB test). The bottom-three methods (Naïve, Holt and Holt-Winters) perform similarly between them but significantly worse than all other methods. The ranked performance of all other methods (Theta, ETS-AutoARIMA, SES, Damped, ETS, SHD and AutoARIMA) is statistically indifferent.

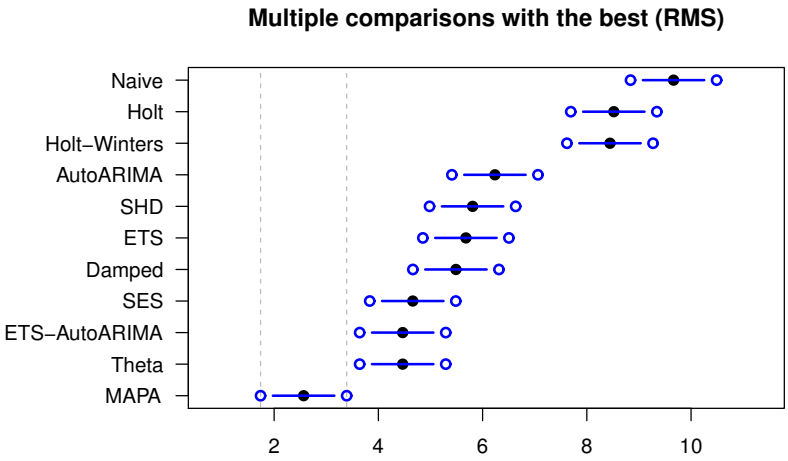


Figure 5: Multiple comparisons with the best based on the $RMS$.

A comparison of Figures 2 and 5 suggests that the differences between methods in terms of forecasting performance are amplified when one is evaluating inventory performance. While methods perform similarly in terms of bias and the top-five methods show statistically indifferent accuracy results, the differences in performance become more distinct with $RMS$. This result agrees with a previous study by Syntetos et al. (2010).

It is worth stretching that the $RMS$ allows for a holistic view of various performance dimensions. For example, the inventory versus achieved service level trade-off curves for lead time 12 (last panel of Figure 4) show that MAPA only achieves only an average performance (it is depicted by the grey color). This result alone contradicts the findings of Kourentzes et al. (2014) who found greater improvements for the longer horizons. However, according to Figure 3, MAPA dominates all other methods and approaches in terms of order and

Petropoulos, F., Wang, X. and Disney, S.M., (2019), The inventory performance of forecasting methods: evidence from the M3-competition data. International Journal of Forecasting, 35 (1), 251-265. DOI: 10.1016/j.ijforecast.2018.01.004.

inventory variance, especially for longer horizons. As such, the $RMS$ allows the integration of different (and potentially contradicting) measures.

One limitation of the RMS as implemented above is that we assumed equal-importance weights for the three utility metrics ($C_h$, $v_o$ and $\alpha$), as equation 13 suggests. However, one could consider unequal weights. We did a sensitivity analysis of the values of RMS using weights for the three metrics so that each weight $\in [0.2, 0.5]$ and the sum of all three weights equals unity. We repeated this 1,000 times, allowing for different combinations of weights. The average RMS values for each case and method were calculated and the differences in the average RMS values for each method are presented as box-plots in Figure 6. It is worth noting that even if there is some small deviation from the equal weights, the relative rankings of the forecasting methods considered in this study remains to a high degree unchanged.

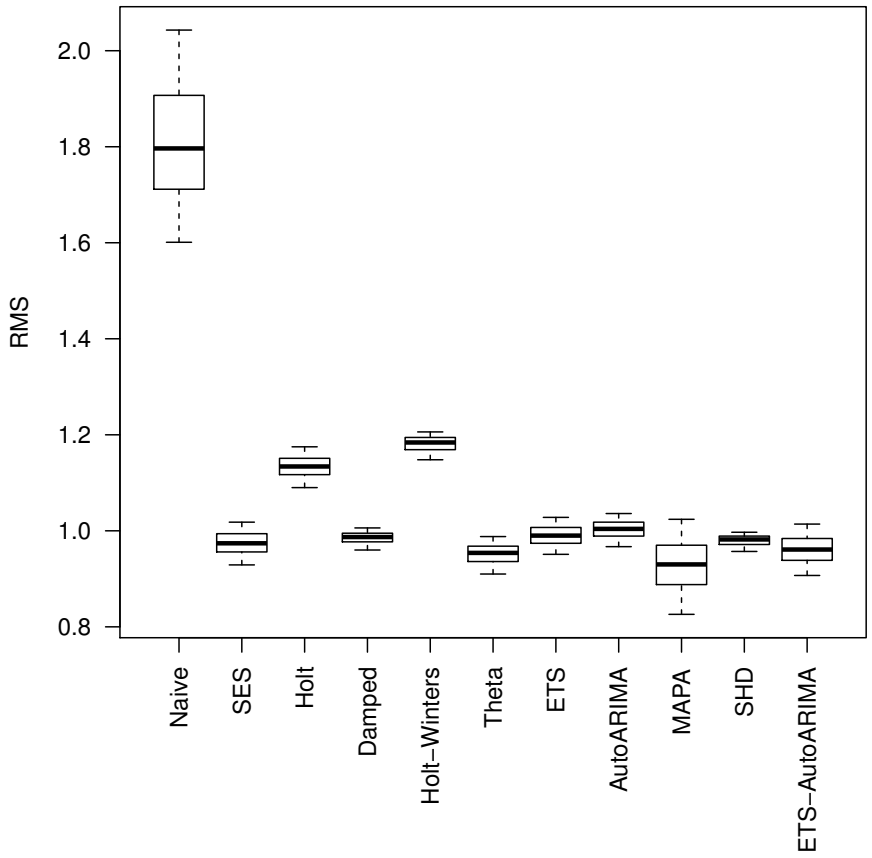


Figure 6: Sensitivity analysis of the average $RMS$ values when different weights for the three utility metrics ($C_h$, $v_o$ and $\alpha$) were applied.

To better understand how the various methods and approaches considered in this study perform under different conditions, we separately consider four categories of the data: series with no trend nor seasonality; series with trend only; series with seasonality only; series with both trend and seasonality. The identification of the time series characteristics was based on the model form of the exponential smoothing model (using the `ets()` function)

Petropoulos, F., Wang, X. and Disney, S.M., (2019), The inventory performance of forecasting methods: evidence from the M3-competition data. International Journal of Forecasting, 35 (1), 251-265. DOI: 10.1016/j.ijforecast.2018.01.004.

that optimally fits the data. To this end, the first five years (60 months) of each series were used. Table 5 presents the series counts in each category.

Table 5: Series counts based on identified characteristics.

| Trend | Seasonality | | Total |
|---|---|---|---|
| | No | Yes | |
| No | 110 | 122 | 232 |
| Yes | 41 | 61 | 102 |
| Total | 151 | 183 | 334 |

Table 6 provides the values for the $RMS$ together with their ranks for each category separately. The respective MCB plots are shown in Figure 7. It should be noted that Table 6 reports arithmetic means of the values of RMS whereas Figure 7 reports the respective mean ranks; this explains some inconsistencies in the ranks for some methods. MAPA is ranked first in three out of four categories in terms of average $RMS$ and also ranked first across all categories in terms of average ranks. Theta and the combination of ETS-AutoARIMA follow. Naïve, Holt-Winters and Holt are the three worst performing methods across all categories, despite the apparent expectation that Holt and Holt-Winters would ranked higher for the trended only and the trended and seasonal series respectively. Overall, we can conclude that the summary results based on the $RMS$ in Table 4 stand if the series are splitted in categories based on their characteristics. One exception is the performance of SES, which as expected is negatively influenced by the existence of trend.

Table 6: Average root mean square ($RMS$) values for the different methods (and ranks) decomposed for data with different time series characteristics.

| Method | No trend, no seasonality | Trend, no seasonality | No trend, seasonality | Trend, seasonality |
|---|---|---|---|---|
| Naïve | 1.733 (11) | 1.697 (11) | 1.844 (11) | 2.005 (11) |
| SES | 0.930 (2) | 1.020 (7) | 0.983 (2) | 1.027 (8) |
| Holt | 1.047 (9) | 1.204 (10) | 1.184 (9) | 1.140 (9) |
| Damped | 0.967 (7) | 0.980 (5) | 1.011 (3) | 0.974 (3) |
| Holt-Winters | 1.160 (10) | 1.134 (9) | 1.219 (10) | 1.185 (10) |
| Theta | 0.930 (2) | 0.910 (2) | 0.975 (1) | 0.983 (5) |
| ETS | 0.953 (5) | 1.040 (8) | 1.019 (6) | 0.985 (6) |
| AutoARIMA | 0.980 (8) | 0.984 (6) | 1.046 (7) | 0.994 (7) |
| MAPA | 0.850 (1) | 0.882 (1) | 1.052 (8) | 0.928 (1) |
| SHD | 0.958 (6) | 0.956 (4) | 1.011 (3) | 0.974 (3) |
| ETS-AutoARIMA | 0.930 (2) | 0.955 (3) | 1.018 (5) | 0.948 (2) |

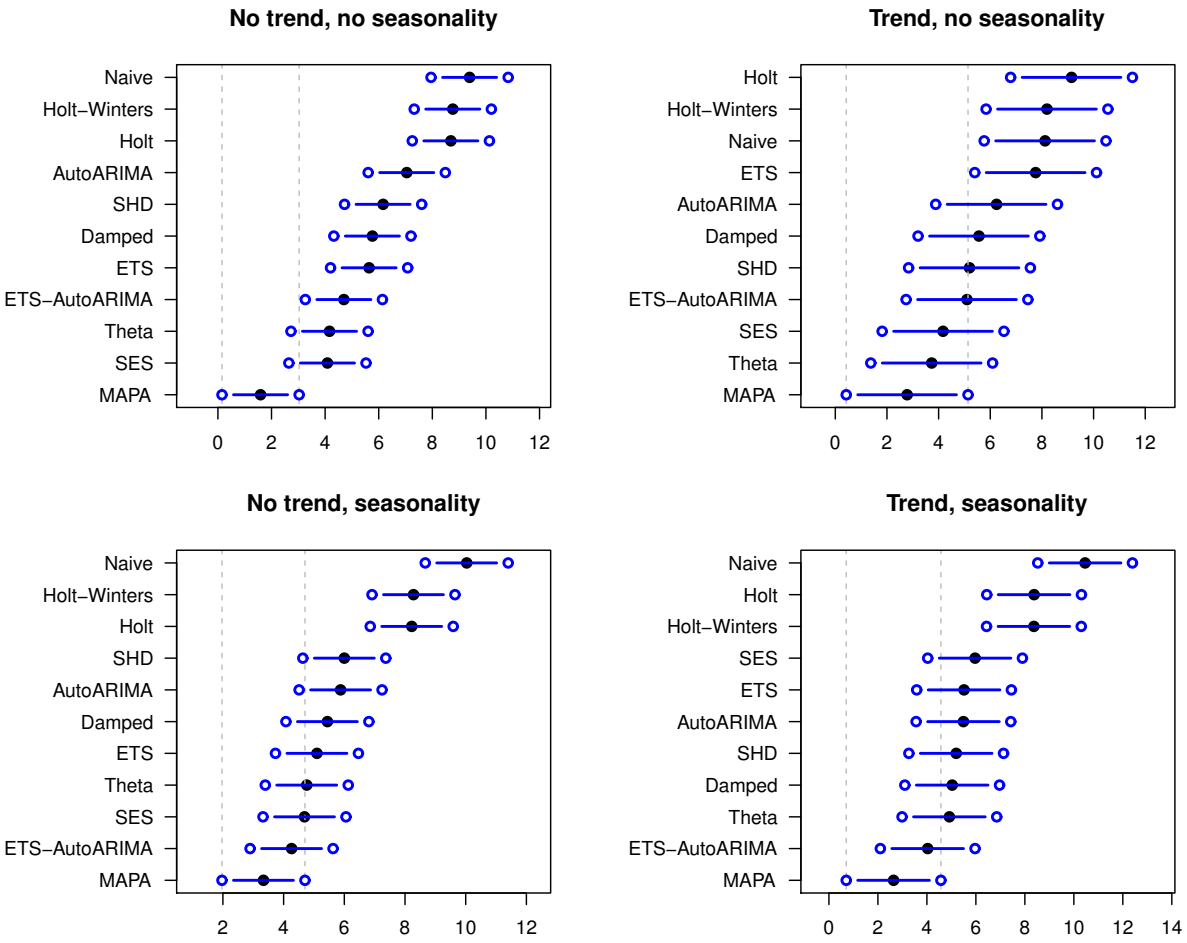


Figure 7: Multiple comparisons with the best based on the $RMS$ decomposed for data with different time series characteristics.

Petropoulos, F., Wang, X. and Disney, S.M., (2019), The inventory performance of forecasting methods: evidence from the M3-competition data. International Journal of Forecasting, 35 (1), 251-265. DOI: 10.1016/j.ijforecast.2018.01.004.

## 6. Conclusion

In this research, we have compared the performance of several popular forecasting methods in terms of inventory control objectives by a simulation approach. The data and several forecasting methods are consistent with those adopted in the M3 competition, making this research an extension to Makridakis and Hibon (2000).

We established an inventory system model based on the order-up-to policy to observe the performance under different scenarios of lead-time and target service levels. We developed an overall measure for the inventory performance which takes into account order variance, inventory holding cost and achieved service level in order to rank these forecasting methods. Below we explicitly address the research questions raised in Section 2.

- *On the utility performance of Holt-Winters method.* Holt-Winters' weak forecasting performance (Makridakis and Hibon, 2000) is coupled with a very weak performance in terms of inventory, being the worst method when inventory cost/service levels trade-offs are considered. Together with Naïve and Holt, are the worst-three methods in this exercise.
- *On the utility performance of exponential smoothing and ARIMA.* Overall, ETS performs better than AutoARIMA, a result that agrees with the relative forecast performance of these two approaches. At the same time, AutoARIMA shows competence for longer lead-times. Also, Damped method's robust forecasting performance is confirmed in terms of inventory, suggesting that it can be considered as a benchmark (Gardner, 1990; Acar and Gardner, 2012).
- *On the utility performance of Theta method.* Theta method is ranked 2$^{\text{nd}}$ overall, confirming its good performance on the M3-competition data. However, its performance is not statistically different to ETS-AutoARIMA or SES.
- *On the utility performance of combinations.* Forecast combinations (SHD and ETS-AutoARIMA) outperform their base methods in terms of inventory performance. Moreover, modern approaches that are based on combinations (MAPA and Theta) are top-ranked.
- *On the utility performance of MAPA.* Multiple temporal aggregation exhibited improved forecasting performance (Kourentzes et al., 2014). At the same time, it produces smooth forecasts minimising the bullwhip effect and the best trade-off curves for inventory cost versus service level. Overall, MAPA results in the best ranked performance as measured by $RMS$, a performance that is significantly different from all other methods.

This study is limited to 334 time series (monthly industry data) from the M3-competition. An obvious path for future research would be to replicate the results presented here using different and richer data sets, both in terms of quantities but also frequencies (weekly or even daily data). Also, in the future we plan to extend this research in terms of different inventory distributions (e.g., arbitrary), service measures (e.g., fill rate) and inventory policies (e.g., $s$-$S$ or $r$-$nQ$). It would also be interesting to investigate forecasting performance via pure economic measures such as the inventory holding/backing and guaranteed hours/overtime approach in Hosoda and Disney (2012).

Petropoulos, F., Wang, X. and Disney, S.M., (2019), The inventory performance of forecasting methods: evidence from the M3-competition data. International Journal of Forecasting, 35 (1), 251-265. DOI: 10.1016/j.ijforecast.2018.01.004.

As a last comment, we would like to note that our empirical study considered a wide range of benchmarks, widely used methods and state-of-the-art approaches coupled with a wide range of performance measures that are relevant in production and operations management. Moreover, we evaluated the performance of the methods over multiple origins (rolling origin evaluation; Tashman, 2000) so that the results are not overfitted on a single evaluation window. We invite future empirical forecasting studies to follow a similar design that allows for rich comparisons over established methods and benchmarks.

## Acknowledgments

We would like to thank the Editor, the Associate Editor and two anonymous reviewers for their constructive comments that helped us improving the original manuscript. We are also grateful to Aris Syntetos and Stephan Kolassa for their insightful comments on earlier versions of the manuscript. Finally, we would like to thank the participants of the International Symposium on Forecasting ISF2016 (June 2016, Santander, Spain) and the IIF workshop on Supply Chain Forecasting for Operations (July 2016, Lancaster, UK) for their comments and suggestions.

Petropoulos, F., Wang, X. and Disney, S.M., (2019), The inventory performance of forecasting methods: evidence from the M3-competition data. International Journal of Forecasting, 35 (1), 251-265. DOI: 10.1016/j.ijforecast.2018.01.004.

## References

Acar, Y., Gardner, Jr., E. S., Oct. 2012. Forecasting method selection in a global supply chain. International Journal of Forecasting 28 (4), 842–848.

Ali, M. M., Boylan, J. E., Syntetos, A. A., Oct. 2012. Forecast errors and inventory performance under forecast information sharing. International Journal of Forecasting 28 (4), 830–841.

Assimakopoulos, V., Nikolopoulos, K., 2000. The Theta model: a decomposition approach to forecasting. International Journal of Forecasting 16 (4), 521–530.

Axsäter, S., 2006. Inventory control. Springer, New York.

Babai, M. Z., Ali, M. M., Nikolopoulos, K., 2012. Impact of temporal aggregation on stock control performance of intermittent demand estimators: Empirical analysis. Omega: The International Journal of Management Science 40 (6), 713–721.

Babai, Z. M., Syntetos, A. A., Teunter, R., 16 Apr. 2010. On the empirical performance of (T, s, S) heuristics. European Journal of Operational Research 202 (2), 466–472.

Barrow, D. K., Kourentzes, N., 2016. Distributions of forecasting errors of forecast combinations: Implications for inventory management. International Journal of Production Economics 177, 24–33.

Blinder, A., Maccini, L., 1991. Taking stock: A critical assessment of recent research on inventories. Journal of Economic Perspectives 5 (1), 73–96.

Boylan, J., Goodwin, P., Mohammadipour, M., Syntetos, A., 2015. Reproducibility in forecasting research. International Journal of Forecasting 31 (1), 79–90.

Boylan, J. E., 2016. Reproducibility. IMA Journal of Management Mathematics 27 (2), 107–108.

Boylan, J. E., Syntetos, A. A., Karakostas, G. C., 2008. Classification for forecasting and stock control: a case study. The Journal of the Operational Research Society 59 (4), 473–481.

Cachon, G., Randall, T., Schmidt, G., 2007. In search of the bullwhip effect. Manufacturing and Service Operations Management 9 (4), 457–479.

Chen, Y. F., Drezner, Z., Ryan, J. K., Simchi-Levi, D., 2000. Quantifing the bullwhip effect in a simple supply chain: The impact of forecasting, leadtimes and information. Management Science 46 (3), 436–443.

Croston, J. D., 1972. Forecasting and stock control for intermittent demands. Operational Research Quarterly (1970-1977) 23 (3), 289–303.

Disney, S. M., Gaalman, G., Hedenstierna, C. P. T., Hosoda, T., 2015. Fill rate in a periodic review order-up-to policy under auto-correlated normally distributed, possibly negative, demand. International Journal of Production Economics 170, 501–512.

Disney, S. M., Towill, D., van de Velde, W., 2004. Variance amplification and the golden ratio in production and inventory control. International Journal of Production Economics 90 (3), 295–309.

Eaves, A. H. C., Kingsman, B. G., 2004. Forecasting for the ordering and Stock-Holding of spare parts. The Journal of the Operational Research Society 55 (4), 431–437.

Fiorucci, J. A., Pellegrini, T. R., Louzada, F., Petropoulos, F., Koehler, A. B., 2016. Models for optimising the theta method and their relationship to state space models. International Journal of Forecasting 32 (4), 1151–1161.

Gardner, E. S., 1990. Evaluating forecast performance in an inventory control system. Management Science 36 (4), 490–499.

Hedenstierna, C., Disney, S. M., 2016. Inventory performance under staggered deliveries and auto-correlated demand. European Journal of Operational Research 249 (3), 1082–1091.

Holt, C. C., Modigliani, F., Muth, J., Simon, H., 1960. Planning production, inventories and the workforce. Prentice-Hall, New Jersey.

Hosoda, T., Disney, S. M., 2006. On variance amplification in a three-echelon supply chain with minimum mean squared error forecasting. Omega: The International Journal of Management Science 34, 344–358.

Hosoda, T., Disney, S. M., 2012. A delayed demand supply chain: Incentives for upstream players. Omega: The International Journal of Management Science 40 (4), 478–487.

Hyndman, R. J., Billah, B., 2003. Unmasking the theta method. International Journal of Forecasting 19 (2), 287–290.

Hyndman, R. J., Khandakar, Y., 7 2008. Automatic time series forecasting: The forecast package for R. Journal of Statistical Software 27 (3), 1–22.

Hyndman, R. J., Koehler, A. B., 2006. Another look at measures of forecast accuracy. International Journal of Forecasting 22, 679–688.

Koning, A. J., Franses, P. H., Hibon, M., Stekler, H. O., 2005. The M3 competition: Statistical tests of the results. International Journal of Forecasting 21 (3), 397–409.

Kourentzes, N., 2013. Intermittent demand forecasts with neural networks. International Journal of Production Economics 143, 198–206.

Kourentzes, N., 2014. On intermittent demand model optimisation and selection. International Journal of Production Economics 156, 180–190.

Kourentzes, N., Petropoulos, F., Trapero, J. R., 2014. Improving forecasting by estimating time series structural components across multiple frequencies. International Journal of Forecasting 30 (2), 291–302.

Lee, H., Padmanabhan, V., Whang, S., 1997. Information distortion in a supply chain: The bullwhip effect. Management Science 43 (4), 546–558.

Li, Q., Disney, S. M., Gaalman, G., 2014. Avoiding the bullwhip effect using damped trend forecasting and the order-up-to replenishment policy. International Journal of Production Economics 149, 3–16.

Liao, W. T., Chang, P. C., Dec. 2010. Impacts of forecast, inventory policy, and lead time on supply chain inventory—a numerical study. International Journal of Production Economics 128 (2), 527–537.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., Winkler, R., 1982. The accuracy of extrapolation (time series) methods: Results of a forecasting competition. Journal of Forecasting 1 (2), 111–153.

Makridakis, S., Hibon, M., 2000. The M3-competition: results, conclusions and implications. International Journal of Forecasting 16 (4), 451–476.

Nikolopoulos, K., Syntetos, A. A., Boylan, J. E., Petropoulos, F., Assimakopoulos, V., 2011. An aggregate - disaggregate intermittent demand approach (ADIDA) to forecasting: An empirical proposition and analysis. Journal of the Operational Research Society 62 (3), 544–554.

Petropoulos, F., Kourentzes, N., Nikolopoulos, K., 2016. Another look at estimators for intermittent demand. International Journal of Production Economics 181, Part A, 154–161.

Sani, B., Kingsman, B. G., 1997. Selecting the best periodic inventory control and demand forecasting methods for low demand items. The Journal of the Operational Research Society 48 (7), 700–713.

Silver, E. A., Robb, D. J., 2008. Some insights regarding the optimal reorder period in periodic review inventory systems. International Journal of Production Economics 112 (1), 354–366.

Snyder, R. D., Koehler, A. B., Ord, J. K., 2002. Forecasting for inventory control with exponential smoothing. International Journal of Forecasting 18 (1), 5–18.

Strijbosch, L. W. G., Syntetos, A. A., Boylan, J. E., Janssen, E., 2011. On the interaction between forecasting and stock control: the case of non-stationary demand. International Journal of Production Economics 133 (1), 470–480.

Syntetos, A. A., Boylan, J. E., 2005. The accuracy of intermittent demand estimates. International Journal of Forecasting 21 (2), 303–314.

Syntetos, A. A., Boylan, J. E., 2006. On the stock control performance of intermittent demand estimators. International Journal of Production Economics 103 (1), 36–47.

Syntetos, A. A., Boylan, J. E., Croston, J. D., 2005. On the categorization of demand patterns. Journal of the Operational Research Society 56 (5), 495–503.

Syntetos, A. A., Nikolopoulos, K., Boylan, J. E., 2010. Judging the judges through accuracy-implication metrics: The case of inventory forecasting. International Journal of Forecasting 26 (1), 134–143.

Syntetos, A. A., Nikolopoulos, K., Boylan, J. E., Fildes, R., Goodwin, P., 2009. The effects of integrating management judgement into intermittent demand forecasts. International Journal of Production Economics 118 (1), 72–81.

Syntetos, A. A., Zied Babai, M., Gardner, Jr., E. S., 2015. Forecasting intermittent inventory demands: simple parametric methods vs. bootstrapping. Journal of Business Research 68 (8), 1746–1752.

Tashman, L. J., 2000. Out-of-sample tests of forecasting accuracy: an analysis and review. International

Journal of Forecasting 16 (4), 437–450.

Teunter, R. H., Duncan, L., 2009. Forecasting intermittent demand: a comparative study. The Journal of the Operational Research Society 60 (3), 321–329.

Wang, X., Disney, S. M., 2016. The bullwhip effect: Progress, trends and directions. European Journal of Operational Research 250 (3), 691–701.

Wang, X., Petropoulos, F., 2016. To select or to combine? The inventory performance of model and expert forecasts. International Journal of Production Research 54 (17), 5271–5282.